Co-funded by the
Erasmus+ Programme
of the European Union

EULALIA

European Latin Linguistic Assessment

Erasmus+ Strategic Partnership for Higher Education (2019-2022)

(2019-1-IT02-KA203-062286)

https://site.unibo.it/eulalia/en

O1: European Latin Language Certification – Basic Level

Methodological and Pedagogical tools

LEXICON

(Italian Version: 31.05.2021)

Project Coordinator:

Alma Mater Studiorum – University of Bologna (Italy)

Project Partners:

University of Köln (Germany)

Catholic University of the Sacred Heart – Milan (Italy)

University of Rouen (France)

University of Salamanca (Spain)

University of Uppsala (Sweden)

Un nuovo “lessico di base” latino

Guido Milanese

1. Stato dell'arte

Il progetto di una valutazione linguistica standard, come progettato da Eulalia, richiede un lessico di base. L'offerta in questo campo è ampia: si vedano ad esempio gli strumenti raffinati offerti dal Progetto Perseus[1] o la lista fornita dal Dickinson College (il team è guidato da Christopher Francese)[2]. Tuttavia, la lista del Dickinson College Dickinson College è limitata a 1000 parole, mentre il progetto Eulalia richiede una maggiore quantità di parole per i livelli più alti. Inoltre, queste liste non informano sulle fonti utilizzate per la compilazione delle liste stesse; ed è stato concordato dai membri del progetto Eulalia di fornire ai potenziali studenti una lista preparata usando testi latini classici standard, cioè i testi che fanno parte del vero "canone" usato nelle scuole europee.

Per questo motivo, il vecchio e venerabile *Vocabulaire de base du latin*, pubblicato nel 1984 da un'équipe di Besançon (A.R.E.L.A.B.) che si è avvalsa del *Dictionnaire fréquentiel* pubblicato dal gruppo LASLA di Liegi nel 1981, può ancora essere usato come una “base” affidabile per costruire un nuovo lessico di base adatto alle esigenze dell’insegnamento del latino nelle scuole e nelle università europee. I vantaggi di questo lavoro sono:

1. quantità di parole: 1600 parole, compatibili con il progetto Eulalia;

2. una lista chiara dei testi latini utilizzati per preparare il lessico:

«Le corpus analysé dans cet ouvrage comprend, en partie ou en totalité, les oevres de Catulle, César, Cicéron, Horace, Juvénal, Ovide, Perse, Properce, Quinte-Curce, Salluste, Sénèque, Tacite, Tibulle, Tite-Live, Virgile, Vitruve»[3].

Nessuna lista è perfetta, naturalmente, e la mancanza di autori come Terenzio o Lucrezio può essere criticabile, ma per preparare una lista "neutra" di parole del latino tardo-repubblicano e di prima età imperiale la scelta degli autori è adeguata. Chi scrive ritiene che il latino medievale possa fornire agli insegnanti una grande quantità di opere molto efficaci per l'insegnamento del latino, ma anche in questo caso il nucleo del lessico può essere facilmente integrato.

2. Scansione e OCR del lessico di Besançon

Il lessico franco-belga è stato pubblicato come libro[4]. Sfortunatamente, la qualità della stampa originale è quella di un libro stampato negli anni '80 con i dispositivi dell'epoca. L'elenco delle parole è stato scansionato utilizzando una HP OfficeJet Pro 7720, e l'OCR è stato eseguito dal noto

*Eulalia – European Latin Linguistic Assessment

[1] https://www.perseus.tufts.edu/hopper/help/vocab

[2] http://dcc.dickinson.edu/vocab/core-vocabulary

[3] G. Cauquil and J.Y. Guillaumin, *Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique)*, ARELAB, Besançon 1984, p. 4.

[4] Lo scarno elenco di parole è stato ristampato in opere più recenti, come Paolo Lamagna, *Il lessico latino di base*, Bompiani, Milano 1999.

programma OCR *tesseract*, con l'opzione *latin*[5]. La lista è stata corretta manualmente ma, con mia sorpresa, anche il primo output di *tesseract* era già molto buono.

3. Etichettatura morfologica

La lista finale fornita dal lessico di Liegi-Besançon è una nuda lista di parole e frequenze, con qualche nota aggiunta qua e là. Al fine di utilizzarla per una vera lista del "lessico di base", era ovviamente necessario aggiungere un tagging morfologico adeguato. A questo scopo, ho usato *treetagger*, il programma morfologico progettato da Helmut Schmid nel progetto TC presso l'Istituto di Linguistica Computazionale dell'Università di Stoccarda e ora ospitato dall'Università di Monaco, Germania[6]. L'esecuzione di *treetagger* su una lista di parole è un approccio soggetto a errori: il parser non può eseguire un'analisi analisi contestuale, e le voci come *quam* sono ovviamente opache[7]. L'analisi morfologica offerta da *treetagger* è stata rivista manualmente; circa il 15% dei tag erano sbagliati, come previsto. Le note di ARELAB sono state utilizzate per correggere l'etichetta di *treetagger*, quando era diversa dalla categoria morfologica offerta dalla nota ARELAB.

4. Il problema delle parole funzione

Durante una riunione del team di Eulalia, Mélanie Lucciano (Université de Rouen – Normandie) ha notato che una lista di frequenza "pura" non era adatta a essere usata come strumento per l'insegnamento di una lingua perché le parole funzione ("mots-outils", "morfemi grammaticali", "parole vuote"...) sono molto comuni in qualsiasi lingua e sono elencate nel rango più alto di una lista di frequenza. Di conseguenza, se supponiamo una lista di 700 parole per un livello A1, gli studenti imparerebbero una grande quantità di congiunzioni e preposizioni, a scapito di verbi, sostantivi e così via.

Le classi di morfologia elencate da *treetagger* sono le seguenti:

- ADJ
- ADJ:NUM
- ADV
- CC
- CS
- DET
- DIMOS

[5] Sul *tesseract* si veda Guido Milanese, *Filologia, letteratura, computer. Idee e strumenti per l'informatica umanistica*, Vita e Pensiero, Milano 2020, pp. 103-107, e la documentazione online su https://github.com/tesseract-ocr. Il programma è FOSS (Free and Open Source Software) ed è stato eseguito su un sistema Linux Mint 20.1.

[6] Si veda https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/, e https://www.cis.lmu.de/~schmid/ per la bibliografia recente.

[7] Non c'è nessuna significativa differenza nell'uso della più recente creazione di Schmid, chiamata "RNNTagger": si veda https://www.cis.lmu.de/~schmid/tools/RNNTagger/.

- ENCL

- ESSE

- EXCL

- FW

- INDEF

- INT

- N

- POSS

- PREP

- PRON

- REL

- V

Dato che si tratta dell'insegnamento della lingua, e lo scopo di adattare l'elenco originale di parole ARELAB è quello di evitare troppe parole funzione, proporrei proporre di considerare come “parole funzione” le seguenti categorie: congiunzioni, preposizioni e clitici.

Un semplice programma SNOBOL4 eseguito sulla lista ARELAB produce questi risultati[8]:

- sull'intera lista (1635 parole) le “parole funzione” sono 78, cioè il 4,1260%

- sulle prime 600 parole, le “parole funzione” sono 55, cioè il 9,100%

- sulle prime 400 parole, le “parole funzione” sono 45, cioè 11,100%

Inoltre, il 57% delle “parole funzione” è elencato tra le prime 400 parole, e quasi il 70% tra le prime 600 parole. La conseguenza è che la lista delle parole più frequenti non è adatta per testare la competenza linguistica degli studenti, perché tra le prime 400/600 posizioni ci sono troppe “parole funzione”. Lo scopo di questa lista rivista è quello di bilanciare le parole funzione e le “parole semantiche”, spostando alcune parole semantiche dalla fascia bassa della lista (= parole dopo n. 400 e n. 600) per ottenere una percentuale più equilibrata di parole semantiche/parole funzione. In altre parole, le prime 400 e 600 parole dovrebbero presentare una percentuale di parole funzione simile a quella dell'elenco completo.

La lista è stata rivista spostando alcune parole funzione nelle posizioni più basse dell'elenco, assicurandosi, tuttavia, che le parole funzione essenziali (come *quia* o *si*) siano elencate nelle liste per i principianti. Il lavoro è stato fatto con un approccio assistito dal computer cioè controllando l'equilibrio delle parole passo dopo passo.

La lista rivista presenta ora le seguenti percentuali:

[8] Sebbene sia un "linguaggio di nicchia", snobol4 è ben mantenuto, in particolare grazie agli sforzi di Phil Budne: si veda https://www.snobol4.org/. La versione più recente è stata resa disponibile nel dicembre 2020. Su questo linguaggio, particolarmente adatto all'analisi dei testi, si veda Milanese, Filologia, letteratura, computer cit., pp. 241-244. Snobol4 funziona su Linux, Windows e OSX.

- sulle prime 600 parole, le "parole di funzione" sono 46, cioè il 7,4%

- sulle prime 400 parole, le "parole funzione" sono 30, cioè il 7,2%.

Anche se queste percentuali non rispecchiano esattamente i dati della lista completa, questo mi sembra un compromesso ragionevole - anche se discutibile, naturalmente, e aperto a ulteriori miglioramenti - tra l'esattezza e le esigenze pratiche di insegnamento.

5. L'output finale

L'output finale è stato prodotto sotto forma di una lista *csv* (Comma Separated Values) adatto ad essere letto da programmi standard come LibreOffice *Calc* o MSOffice *Excel*. Ho aggiunto una semplice intestazione:

N, VOX, MORPHO, NOTAE, FREQ

dove **N** = numero, **VOX** = la parola elencata, **MORPHO** = la frase morfologica, **NOTAE** = la nota originale del Vocabulaire Liège-Besançon, e **FREQ** la frequenza della parola. Per esempio:

16, QUIS, PRON, interr., 4555

L'informazione aggiunta dal *Vocabulaire* serve a specificare il tipo di pronome al quale *quis* appartiene. Ogni volta che l'informazione fornita dal *Vocabulaire* è la stessa del tag morfologico, essa viene silenziosamente omessa. Per rendere la lista più leggibile, è stato prodotto un file PDF dalla lista CSV, utilizzando *pandoc*[9].

6. Elenchi dalla lista

Dato che lo scopo di tutto questo è fornire agli insegnanti e agli studenti strumenti comodi e facili da usare, facili da usare, ho riorganizzato le liste in varie forme. Il primo e ovvio passo è stato quello di stampare le liste delle prime 400 e 600 parole.

6.1 Liste alfabetiche

Tutte le liste (cioè l'intera lista e le liste "400" e "600") sono state ordinate in ordine alfabetico, in modo che l'insegnante o lo studente possa individuare una data parola sfogliando la lista in ordine alfabetico. La lista alfabetica completa è stata anche trasformata in un file PDF.

6.2 Liste morfologiche

Tutte le liste (cioè l'intera lista e le liste "400" e "600") sono state ordinate in ordine morfologico, in modo che l'insegnante o lo studente possano individuare tutte le parole di una data categoria morfologica. La lista morfologica completa è stata anche trasformata in un file PDF.

[9] Si veda https://pandoc.org.

7. Una GUI della lista

Un'interfaccia grafica molto semplice, per facilitare l'uso di queste liste, è stata sviluppata da chi scrive ed è disponibile per i beta tester interessati a questo progetto.

8. Codice degli script snobol4 e bash

8.1 Lo script snobol4

Lo script snobol4 per il calcolo delle percentuali:

```
        P_type = break(',') len(1) break(',') len(1) (break(','). M_type)
        KL =0
        KE =0
loop    Line = input    :f(loop_end)
        KL = KL +1
        Line ? P_type
        ((leq(M_type,"CC")), (leq(M_type,"CS")), (leq(M_type,"ENCL")),
+       (leq(M_type,"PREP"))) (KE = KE +1) :(loop)
loop_end        Output = KE'/' KL
        Extreme = KE *100
        Percentage = Extreme / KL
        Modulus = Remdr(Extreme,KL)
        Terminal = Percentage'.' Modulus
end
```

La prima linea dichiara un *pattern*: in una data linea CSV, il cursore si sposta alla prima virgola, poi di un carattere, ripete gli stessi passi, e al 3° tentativo salva l'informazione come tipo morfologico. Poi inizializziamo il contatore di linee (KL) e il contatore delle parole vuote (KE) a zero[10]. Il ciclo legge tutte le righe dallo standard output; se il tipo morfologico è una parola vuota, il contatore KE viene aumentato di 1. Dopo il ciclo, alcune semplici operazioni aritmetiche producono l'output desiderato.

Lo stesso programma viene eseguito sulle prime 400 / 600 parole aggiungendo un limite:

```
lt(KL,601)
```

In questo caso, il ciclo terminerà alla 600a riga (lt = meno di).

8.2 Gli script bash

Gli script bash usati per disporre le liste in ordine alfabetico o morfologico sono semplici script di ordinamento, come il seguente:

```
sort -t, -k3,3 -k2 frequentia-rev-1-400.csv\
> frequentia-rev-1-400-morph.csv
```

[10] Questo passo non è necessario, ma l'inizializzazione dei dati rende il programma più facile da leggere e da capire.

Questa istruzione ordina la lista "400" sul 3° campo (il tipo morfologico) e poi sul 2° (= in ordine alfabetico), scrivendo l'output in un nuovo file (*frequentia-rev-1-400-morph.csv*). Una virgola è usata per separare i campi (“-t,” o “-t','”).

8.3 Da CSV a PDF usando pandoc

Pandoc è un programma a riga di comando molto efficiente in grado di tradurre da molti formati attualmente utilizzati (ad esempio DOCX, LATEX, markdown) in altri formati[11]. La più recente versione di Pandoc può leggere un file CSV e trasformarlo in altri formati, incluso PDF (un silenzioso richiamo a TEX).

Questa è la semplice istruzione:

```
pandoc -f csv frequentia-format-rev-morph.csv\
-s -o frequentia-format-rev-morph.pdf
```

Pandoc legge un file CSV (-f significa "da") e produce un file PDF. Lo stesso schema si applica a tutti gli altri file.

Bibliografia

- Cauquil, G. and J.Y. Guillaumin, *Vocabulaire de base du latin (alphabétique, fréquentiel, étymologique)*, ARELAB, Besançon 1984.
- Delatte, L., Et. Evrard, S. Govaerts, and J. Denooz, *Dictionnaire fréquentiel et index inverse de la langue latine*, L.A.S.L.A. (Laboratoire d’analyse statistique des langues anciennes),
- Université de Liège, Liège 1981.
- Lamagna, Paolo, *Il lessico latino di base*, Bompiani, Milano 1999.
- Milanese, Guido, *Filologia, letteratura, computer. Idee e strumenti per l’informatica umanistica*, Vita e Pensiero, Milano 2020.

[11] Si veda https://pandoc.org/.